ULTRA-LOW-POWER VOICE-ACTIVITY-DETECTOR THROUGH CONTEXT- AND RESOURCE-COST-AWARE FEATURE SELECTION IN DECISION TREES

Steven Lauwereins[1], Wannes Meert[2], Jort Gemmeke[3], Marian Verhelst[1]

[1]ESAT-MICAS KU Leuven, Belgium
[2]CS-DTAI KU Leuven, Belgium
[3]ESAT-PSI KU Leuven, Belgium

ABSTRACT

Voice-activity-detectors (VADs) are an efficient way to reduce unimportant audio data and are therefore a crucial step towards energy-efficient ubiquitous sensor networks. Current VADs, however, use computationally expensive feature extraction and model building algorithms with too high power requirements to be integrated in low-power sensor nodes. To drastically reduce the VAD power consumption, this paper introduces a decision tree based VAD with (1) a two-phase VAD operation to maximally reduce the power-hungry learning phase, (2) a scalable analog feature extraction block, and (3) context- and dynamic resource-cost-aware feature selection. Evaluation of the VAD was performed with the NOIZEUS database, demonstrating a comparable performance to SoA VADs such as Sohn and Ramírez, while reducing the feature extraction power consumption up to approximately 200 fold.

Index Terms— cost-aware VAD, context-aware machine learning, low-power sensor interface, adaptive circuits

1. INTRODUCTION

Interest in ubiquitous sensor networks is again strongly increasing, spearheading applications relying on smart objects and smart environments. The sensors in these networks are expected to operate autonomously and continuously throughout their complete lifetime of multiple years. This restricts the power budget of such sensors to a few μW while active, which is difficult to attain for current sensors [1]. It is therefore important to discard irrelevant data as early as possible, in order not to waste scarce computational resources on the incoming sea of data. In acoustic sensing, a voice-activity-detector (VAD) is a very efficient way to reduce unimportant audio data, and is often used in speech recognition, speech coding and speech enhancement.

Generally speaking, VADs extract acoustic features from audio and then build discriminative models on those features to classify the input as speech versus non-speech. Early algorithms used simple energy-based features to perform the classification. To increase accuracy, later VADs used more complex features such as zero-crossing rate [2] or pitch [3]. Nowadays, most VADs employ statistical models and use more complex classification models that allow for a more accurate distinction between speech and non-speech. However, this improved accuracy came at the expense of severely increased computational complexity in feature extraction and model building. Ying et al. [4] have acknowledged this problem and optimized their sequential GMM (sGMM) based VAD algorithm to increase its computational efficiency. The computational complexity of the model updating algorithm was reduced by an estimated factor of 100, through re-estimating the sGMM model only with the newest data point instead of the previous $M + 1$ data points. Yet, the feature extraction part still relied on a power-hungry 256pt FFT. A system-level exploration of this VAD indicates the power consumption of its hardware implementation to be at least $150\mu W$, including $10\mu W$ front-end power, $40\mu W$ ADC power and $100\mu W$ processing power in $90nm$ CMOS. This is still a factor 10 too high to be implemented in self-sustainable sensor node applications.

Our previous work [5] was aimed at reducing the total VAD system power through context- and feature-cost-aware feature selection only. This paper extends this work, targeting significantly improved savings, while maintaining comparable accuracy with state-of-the-art (SoA) VADs. This is achieved through the combination of three key innovations:

1. Two-phase operation: a sporadic compute intensive learning phase and a continuous low-complexity inference and quality control phase.
2. Low-complexity scalable energy-based feature extraction in the analog domain, enabling mutually independent feature (de-)activation.
3. Context-aware and dynamic resource-cost-aware feature selection through machine learning methods in the VAD, allowing the system to dynamically only activate low-cost context-relevant features.

To clarify this strategy, Section 2 introduces the overall system architecture of the proposed VAD and discusses the first two innovations. Section 3 explains the third innovation, i.e. the adaptive and power-efficient feature selection method and the algorithm developed to optimally use the introduced hard-

ware adaptivity. Section 4 describes the experimental setup used to evaluate the VAD and summarizes the results and performance of the proposed approach.

2. PROPOSED VAD ARCHITECTURE

A VAD is composed out of three elements, a feature extraction (FE) element, a model building element and a classifier. Algorithm decomposition of SoA VADS [6][7] show that the main power consumers are the FE and modeling. Our proposed VAD (Fig. 1) introduces a novel architecture implementing a two-phase operation decision tree VAD and scalable analog FE, reducing the power of all three elements.

2.1. Two-phase VAD operation

The proposed VAD operates in two distinct phases: a learning phase and an inference phase.

The first phase activates all blocks of Fig. 1 except the decision tree (DT) classifier and the classification quality monitor. A digital-signal-processor (DSP) performs a compute intensive unsupervised classification of the incoming data, making use of a SoA VAD algorithm. Subsequently, a DT is learned on the training set, which is labeled by the unsupervised classifier. This results in a DT that is learned on a specific but not predefined context. In our case, context is determined by the nature of the background noise. The resulting high power consumption of the learning phase is allowable since the phase will only be sporadically activated and thus reduces the model building power consumption. It is activated when the classification quality has degraded below an acceptable level, suggesting a context switch.

The second phase activates only the passive microphone, the analog feature extractor, a slow analog-to-digital-converter (ADC), dedicated digital classifier hardware and the classification monitor. The overall power consumption of this phase is low since all these blocks have low complexity and are dedicated ultra-low-power blocks. The inference phase runs until the classification monitor detects insufficient classification quality and then re-activates the learning phase.

2.2. Feature Extraction

The high power consumption of FE in current VADs is caused by the high mathematical complexity of the digitally calculated features (e.g. DFT, cross-correlation) and by the need for a high speed ADC converting the raw analog audio to the digital environment. We propose to reduce this power usage by extracting power-friendly energy-based features directly in the analog domain (Analog FE block in Fig. 1), reducing the required ADC speed and eliminating the digital calculations computed on the DSP, previously needed for FE. Every feature is extracted by activating a limited number of analog resources. Though the introduced feature-selection paradigm is applicable on any feature set, this paper uses as features the energy difference between neighboring mel-shaped frequency bands, thanks to its straightforward analog implementation:

$$x(k) = E(k) - E(k-1) \tag{1}$$

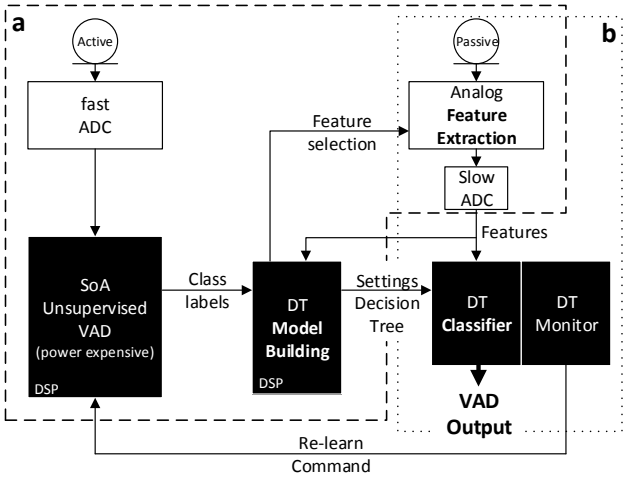

Fig. 1: System architecture of the proposed decision tree (DT) VAD, where white blocks are implemented in analog and black blocks are digitally implemented. Block **(a)** is activated during learning phase, block **(b)** is activated during inference phase.

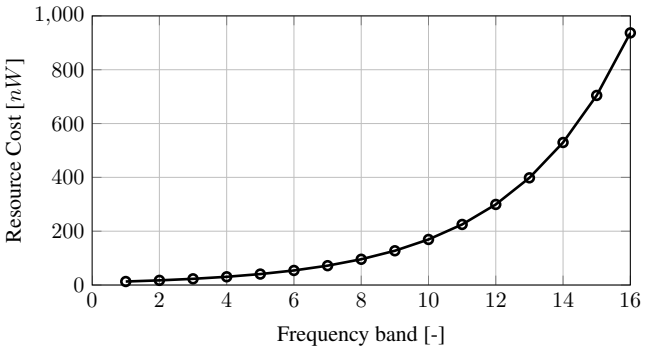

Fig. 2: The power consumption of the analog resources. Features can use multiple resources but from a power point of view preferably as few as possible.

where $x(k)$ denotes feature k of N available features and $E(k)$ the sensed energy in frequency band k, over a time frame extracted by an analog resource:

$$E(k) = \overline{|butter_{(f_l(k),f_h(k))}(y)|} \tag{2}$$

with $butter$ a first order butterworth filter with frequency limits $f_l(k) = 30Hz \cdot 1.33^{k-1}$, $f_h(k) = 30Hz \cdot 1.33^{k}$, $k \in [1; N]$ and y the amplified signal of the passive microphone in a frame. Every energy value $E(k)$ is independently extracted in the analog domain through an analog filter stage, which we will denote as a basic analog resource with a particular power consumption cost.

The analog FE hardware is implemented such that all N analog resources can be shut down individually. As such, reducing the number of used features during classification allows to de-activate corresponding analog resources and hence strongly impacts the system power consumption. This in-

creases the overall FE efficiency and allows us to increase its energy scalability. A challenge to implement such strategy stems from the non-constant power consumption of the analog filter resources, which is frequency dependent (see Fig. 2). Additionally, the nature of the selected feature causes analog resources to be shared by multiple features, further complicating optimal feature selection.

2.3. Model building

The high power consumption of model building in current VADs stems from the high update frequency of the voice/noise model. This enables the VAD to track changes in the voice and noise levels caused by a changing environment and thus increasing their classification accuracy. Switches between noise environments however only occur sporadically, expected at most once per minute. This work therefore proposes to detect context switches by tracking the quality of the classification and to only update the used model once the classification quality drops below a set threshold. This results in longer sleep times for the DSP, which executes the model learning, and thus in a decreased power consumption. The quality of classification can for example be monitored by counting how often the tree enters a fuzzy leaf, i.e. a leaf that during training classified a similar amount of speech and non-speech frames. Alternatively, the quality can be monitored by sporadically activating the DSP classifier and comparing both classifications. The effect of this context switch detection on the power consumption of the VAD is target application specific and beyond the scope of this paper. It will be explored in more detail in future work.

2.4. Classifier

Most VADs classify the incoming data with computationally expensive methods such as a likelihood ratio test in combination with statistical models such as GMMs [6]. Others use low-complexity thresholding of the data to make distinction between speech and non-speech [4]. It is clear from a power perspective, that the second approach is preferable over the first, if sufficient accuracy can be guaranteed. The classifier used in the proposed VAD uses a low-complexity thresholding classifier acting on multiple features, namely a DT [8]. One of the advantages of DTs over other classifiers is that it is efficiently implementable in a dedicated ultra-low-power classification block, further reducing the power consumption. Another advantage is the transparency of DTs with regard to feature importance and usage, enabling context- and dynamic resource-cost-aware feature selection. Since DTs are supervised classifiers, the classifier will be assisted during the learning phase by an unsupervised classifier that labels the training data. This classifier implements a SoA, yet power-inefficient unsupervised VAD such as proposed by Ramírez [7]. This principle of two-stage learning where the first step uses a complex but powerful classifier and the second stage uses the classification information from the first step to learn a more compact and efficient classifier is called model compression [9]. Model compression recently gained importance and has been applied successfully in many domains to significantly reduce the model complexity while maintaining accuracy.

3. CONTEXT- AND COST-AWARE DECISION TREE

This paper proposes to reduce VAD power consumption through the introduction of three elements: two-phase operation, scalable analog FE, and context-aware and dynamic resource-cost-aware feature selection, of which the first two are discussed in Section 2. This section enables context-aware and dynamic resource-cost-aware feature selection through algorithmic changes of DT learning.

3.1. Scalable operation

The introduction of an analog scalable FE block in section 2, where features can be independently (de-)activated, allows for two complementary power-saving feature selection strategies: context-aware and dynamic resource-cost-aware feature selection.

Firstly, the relative information content of a feature is highly context specific. In our case, context is determined by the nature of the background noise. Therefore, only the discriminative features within the current operating context are activated, thus dynamically modifying the amount of required features and active hardware resources. As previously shown [5], context-aware feature selection reduces the power consumption of the feature extraction block by disabling the hardware resources of the non-discriminative features.

Secondly, as discussed in Section 2 and Fig. 2, not every analog resource consumes the same amount of power, and resources are shared between neighboring features. The added power consumption caused by selecting an additional feature is therefore dependent on the resources that are already activated due to other selected features. To optimize the power consumption of the VAD it is crucial to dynamically take the feature-specific or even better the resource-specific power-usage into account during DT building. In DTs this can be done by including the resource-specific power-usage in the cost-function of the DT learning algorithm. The cost-function therefore becomes *information gain* / *Watt* and will be detailed in next subsection.

3.2. Algorithmic implementation

The most commonly used algorithm to learn decision trees is C4.5 [8]. The introduction of context-aware feature scalability is straightforward, as C4.5 already selects the most discriminative features recursively to build its tree. Learning a DT on a specific context therefore allows for context-aware feature selection, which after tree pruning guarantees the activation of only the most relevant analog resources and unused feature de-activation. However, the extension of C4.5 to cost-awareness requires some changes in the used cost-function, to ensure sufficient bias towards selection of features with a low added power-cost. The proposed algorithm is therefore

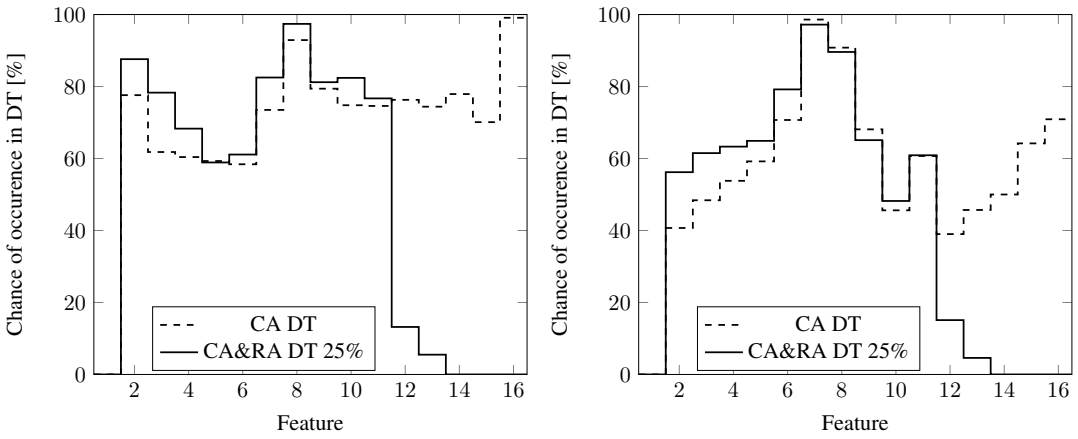

Fig. 3: Chance of occurrence of a feature in a learned DT for context-aware (CA) DTs, and context- and dynamic resource-cost-aware(CA&RA) DTs with the analog FE block only using 25% of its maximal power consumption under babble noise (a) and exhibition noise (b). CA&RA DTs use less expensive features than CA DTs, thus reducing the FE power consumption.

based on a modified cost function, which jointly considers a feature's discriminative nature, with its added resource specific power-cost. The cost-function which is described by a split criterion that has to be maximized becomes:

$$Split_{(n,k)} = \frac{IG_{(n,k)}}{(1-\alpha)\cdot \delta P_{(n,k)} + \alpha \cdot P_{(n-1)}} \qquad (3)$$

with $IG_{(n,k)}$ the information gain [8] of feature k at build step n, $\delta P_{(n,k)}$ the additional power consumption caused by the addition of feature k in build step n taking all resources into account that were activated by feature selection in the $n-1$ previous build steps, $P_{(n-1)}$ the total power consumption of the built tree up to build step $(n-1)$ and a weight factor α. Equation (3) encourages re-use of features, because $\delta P_{(n,k)} = 0$ when feature k was already used higher up in the tree. The additional power consumption $\delta P_{(n,k)}$ for a feature k at step n reusing already activated hardware (already activated analog resources from the analog FE block) will also be smaller than when non of its required resources were already activated. This formulation of cost-aware feature selection can thus be called *dynamic resource-cost-aware* feature selection instead of *feature-cost-aware* feature selection. The formulation of this equation stresses power consumption more in the beginning of the tree construction, when $P_{(n-1)}$ is still small. This ensures cheaper features at the top of the tree which improves the scalability of the total power consumption. As seen in Fig. 3 context-aware and dynamic resource-cost-aware feature selection prioritizes power inexpensive features over their power expensive counterparts, in our case preferring low frequency features.

4. PERFORMANCE EVALUATION

In this section, we evaluate the proposed VAD's performance. It is compared with other leading VADs on the NOIZEUS-database [10]. The proposed VAD is referred to as DT in the following sections.

4.1. Experimental conditions

To assess the performance of the proposed VAD and compare it with the SoA, the simulations are performed with voice and noise files from the NOIZEUS-database [10], the voice files exist out of 30 sentences of 2 seconds each. A ground-truth is achieved by running the VAD of Ramírez [7] over the noise free voice files. This is done for DT training purposes without hangover and for testing with a hangover of $120ms$. Equation (1) describes the features and equation (2) the resources used in the DT simulations. Table 1 shows the other specifics of the experimental setup for the DTs. The ROC curves of the DTs are made by varying the penalty for speech misclassification from 1 to 3 and for non-speech from 1 to 30 and every simulation point is the average over 1000 DTs. To achieve a range of SNRs for training and testing, the amplitude of the voice files is scaled with regard to the average noise and voice power. For SoA comparison, the test audio is also classified by VADs of Sohn [6] and Ramírez [7] with a frame shift of $10ms$, frame length of $20ms$ and hangovers of $120ms$. The ROC curves of the baseline are achieved by varying the voice threshold for both Ramírez and Sohn from 0.0 to 1.0.

4.2. Performance comparison

We designed two experiments to evaluate the discrimination capability of the DT VAD.

Table 1: Values of parameters used in the implementation of the proposed algorithm, for a sampling rate of $8kHz$

frame shift: 10ms	frame length: 50ms	N:16
hangover: 120ms	training SNR: 0-5dB	α : 0.75
#training frames: 500	#test frames: 15930	

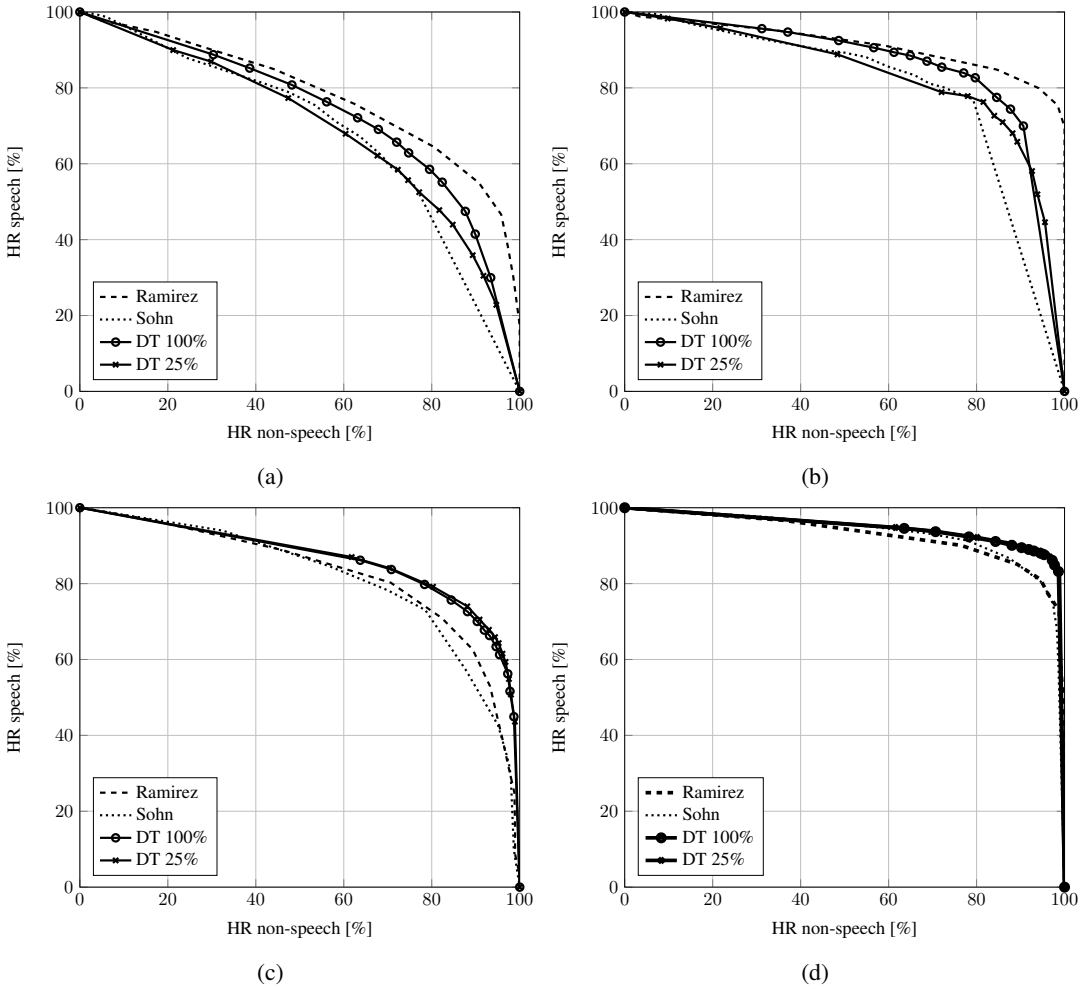

Fig. 4: ROC curves under different noises (rows) and SNRs (columns). (a) 0-dB babble noise. (b) 10-dB babble noise. (c) 0-dB exhibition noise. (d) 10-dB exhibition noise.

Table 2: Accuracy versus maximum allowed FE cost at 5dB SNR. Ramírez and Sohn have a fixed feature extraction cost a factor 38 and 24 higher than the maximal power consumption of the here proposed analog FE block.

Classifier name	Power	Accuracy	
		Babble	Exhibition
Ramírez	$143\mu W$	81%	83%
Sohn	$90\mu W$	72%	82%
100% DT	$3.74\mu W$	77%	88%
50% DT	$1.87\mu W$	75%	88%
25% DT	$0.937\mu W$	75%	88%
10% DT	$0.374\mu W$	72%	88%

The first experiment compares the VAD performances of two DT VADs with the baseline VADs for different noise contexts. Fig 4 shows the ROC curves for babble and exhibition noise at 0 and 10dB SNR. The first DT VAD referred to as DT 100%, is trained to use all features and therefore consumes maximal power, being $3736nW$. The FE extraction block of the second DT is trained to maximally use 25% of this power consumption, or equivalently $935nW$, referred to as the 25% DT. This is achieved by learning the tree breadth-first until the maximum allowed tree cost is reached or until the *information gain / Watt* drops below a threshold, and hereafter pruning the built tree to reduce over-fitting. For babble noise, the 100% DT performance lies between the performance of Sohn and Ramírez, while the 25% DT performs equally good as Sohn over the whole tested SNR range. With exhibition noise both DT VADs outperform Ramírez and Sohn. The DTs lose some accuracy in babble noise because of the signal's time-varying nature, demonstrating less white noise like behavior than exhibition noise.

The second experiment investigates the classification performance of the proposed DT approach in function of its maximum FE power consumption. This power is displayed in both percentages of the maximal power and the actual power consumption of the FE block, as we designed it on $90nm$

CMOS. Table 2 shows that for babble noise the accuracy of the DT slightly increases with increasing power consumption, while for exhibition noise the 10% DT is already at the maximum accuracy. Inspection of the trees built with exhibition noise indicate that this noise is best modeled with the low-frequency features (Fig. 3(b)), which are also the cheapest features explaining its flat accuracy versus power curve, Table 2. Inspection of trees built with babble noise on the other hand show that this type of noise requires high-frequency and thus expensive features (Fig. 3(a)). This allows for a trade-off between power consumption and accuracy, which is useful for sensor networks. The sensors in these networks can dynamically decide to reduce their accuracy when operating in power scarceness, or they can increase their accuracy whenever required, sacrificing power.

We estimate the power consumption of the resources (frequency bins of the raw audio) in the VADs of Ramírez and Sohn to be $143\mu W$ and $90\mu W$ respectively. This estimates $10\mu W$ for amplification of the microphone, $40\mu W$ for an ADC running at $8kHz$ with $16b$ precision [11] and for Ramírez a $512pt$ FFT and Sohn a $256pt$ FFT at $100Hz$ giving $93\mu W$ and $40\mu W$ respectively [12]. These power numbers include for Ramírez and Sohn all hardware required for comparable functionality of the Analog FE block in Fig. 1. Table 2 also shows that the fully activated analog FE block gains a power reduction of a factor 38 compared to Ramírez and 24 compared to Sohn. The context-aware and dynamic resource-cost-aware feature selection gains another factor 10 with a small accuracy penalty. The feature extraction power gain compared to the benchmark SotA VAD's therefore lies between a factor of 24 and 240 compared with Sohn and between 38 and 380 compared with Ramírez. The complete power gain is even larger since the classifiers of Ramírez and Sohn require cross correlations of their features and other computationally complex calculations, which is not included in the power numbers of Table 2. This leads to significantly higher power consumption than the thresholding needed in the here proposed VAD.

5. CONCLUSION & FUTURE WORK

In this paper, we proposed a new VAD architecture enabling VAD usage in sensor networks. This framework outperforms the power consumption of conventional VADs because of two-phase operation, flexible usage of a scalable analog feature extraction block, and context- and dynamic resource-cost-aware feature selection.

The two-phase operation allows the system to work in its most power efficient phase (the inference phase) until its performance drops below acceptance and relearning is required. Analog feature extraction enables extreme power scalability and reduces the overall power consumption by reducing the sample rate of the ADC and the amount of calculations to be computed on a DSP. Context- and dynamic resource-cost-aware feature selection further decreases the power consumption of the VAD by smartly selecting only those features that carry the highest information relative to their power requirements. The proposed VAD has a speech/non-speech accuracy comparable to existing SoA VADs while reducing the required power for feature extraction by a factor 24 to 380. This VAD is just one application of the context- and dynamic resource-cost-aware model compression framework.

The overall power consumption is defined by a trade-off between the low power VAD and the sporadic classification-quality-control by the SoA unsupervised VAD. The optimal point depends on the specifications of the target application and will be explored in future work.

Acknowledgments

This research was funded by FWO-Vlaanderen.

6. REFERENCES

[1] Nick Van Helleputte, "18.3 A multi-parameter signal-acquisition SoC for connected personal health applications," in *International Solid-State Circuits Conference*, 2014, pp. 314–316.

[2] ITU, "Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear prediction (CS-ACELP). Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70,," in *International Telecommunication Union*, 1996.

[3] ETSI, "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, vol. 2, 1999.

[4] Dongwen Ying, Yonghong Yan, Jianwu Dang, and F K Soong, "Voice Activity Detection Based on an Unsupervised Learning Framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2624–2633, 2011.

[5] Steven Lauwereins, Komail Badami, Wannes Meert, and Marian Verhelst, "Context- and cost-aware feature selection in ultra-low-power sensor interfaces," in *ESANN*, 2014, pp. 93–98.

[6] Jongseo Sohn, NS Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1998–2000, 1999.

[7] Javier Ramírez and JM Górriz, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," *Signal Processing Letters, IEEE*, vol. 13, no. 8, pp. 497–500, 2006.

[8] JR Quinlan, *C4.5: Programs for Machine Learning.*, vol. 4, MORGAN KAUFMAN PUBL Incorporated, 1993.

[9] LJ Ba and R Caurana, "Do Deep Nets Really Need to be Deep?," *arXiv:1312.6184*, pp. 1–6, 2013.

[10] Yi Hu and Philipos C Loizou, "Subjective comparison and evaluation of speech enhancement algorithms.," *Speech communication*, vol. 49, no. 7, pp. 588–601, July 2007.

[11] B. Murmann, "ADC Performance Survey 1997-2014," Tech. Rep., Stanford University, 2014.

[12] Texas Instruments, "FFT Implementation on the TMS320VC5505, TMS320C5505, and TMS320C5515 DSPs," Tech. Rep., 2013.